Two-Phase Clustering Strategy for Gene Expression Data Sets

Dirk Habich, Thomas Wächter, and Wolfgang Lehner
Dresden University of Technology
Database Technology Group
dbgroup@mail.inf.tu-dresden.de

Christian Pilarsky
Dresden University of Technology
Visceral Thorax Surgery Group
christian.pilarsky@mailbox.tu-dresden.de

Abstract

In the context of genome research, the method of gene expression analysis has been used for several years. Related microarray experiments are conducted all over the world, and consequently, a vast amount of microarray data sets are produced. Having access to this variety of repositories, researchers would like to incorporate this data in their analyses to increase the statistical significance of their results. In this paper, we present a new two-phase clustering strategy which is based on the combination of local clustering results to obtain a global clustering. The advantage of such a technique is that each microarray data set can be normalized and clustered separately. The set of different relevant local clustering results is then used to calculate the global clustering result. Furthermore, we present an approach based on technical as well as biological quality measures to determine weighting factors for quantifying the local results proportion within the global result. The better the attested quality of the local results, the stronger their impact on the global result.

1 Introduction

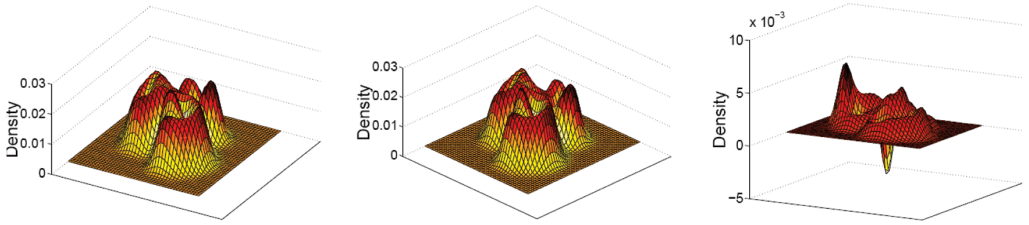
Deoxyribonucleic acid (DNA) microarrays are an important part of a new and promising field of biotechnology. They allow the simultaneous measurement of expression values in cells for thousands of genes. Microarray experiments are increasingly popular in biological as well as medical research to address a wide range of problems. One prominent example is cancer research, where microarrays are used to study the molecular variations among tumors with the aim of developing better diagnostics and treatment strategies. Within the last few years, a vast amount of microarray data sets has been produced for various studies worldwide and made commonly available in public data repositories. Having access to this variety of repositories, researchers would like to incorporate this data sets in their analyses to increase the statistical significance of their results.

The classic gene expression analysis process consists of four steps: *data integration*, *data normalization*, *data analysis* and *interpretation*. The order of the four steps is typically fixed; however, the algorithms within every step are very flexible and not standardized. In this workflow, many different tools are currently involved, which are working in an independent and incompatible way. A first challenge in the *data integration* step is to locate relevant data sets, to download them, and finally, to build up an integrated data base. This step usually requires a lot of time. Due to variations in the experimental conditions and the quality of the biological material, the measurements are not directly comparable and appropriate normalization has to be applied. As the chosen normalization has a strong influence on the analysis results [7], it is desirable to adjust the normalization method according to a data set's characteristics and separate for all the respective data sets.

By following the classic approach for meta-analysis, a combined global data set is normalized to achieve comparability of independently measured expression levels leading to a heterogeneous view of the data, that is not necessarily corresponding to the biological truth. Furthermore, it can be said that data it analyzed in a conjoint manner, which underlies a huge experimental variance. The last step of the classic analysis approach is the analysis of a global normalized data set. As an inherently data-driven technique, clustering can determine the statistical characterization of unknown data distribution, whereas clustering results depend on the underlying data characteristics as well as the initial parameters of the algorithm. Therefore, different clustering results can be produced from one single data set. It is also desirable to adjust the clustering method as well as the normalization according to each single microarray data set and to compute a global result afterwards.

We suggest that each microarray data set should be normalized and clustered separately (first phase) and that the combination of the local clustering results to a global result (second phase) yields better results than the classical meta-analysis. Summarizing the advantages of our two-phase clustering strategy in comparison to the classic approach, it can be said:

- The new approach possibly leads to better results by evaluating single homogeneous microarray data sets instead of just one fully integrated data base, because normalization and clustering can be adjusted for each data set separately.
- The global result is calculated using a set of different local clustering results. In this case, we integrate results instead of data and then analyze the integrated data.
- For every local result, a statistical weighting factor can be determined for quantifying the local results proportion within a global result based on technical and biological quality measures. The better the attested quality of the local results, the stronger their impact on the global result.



(a) $\hat{f}^D$, $n = 200$ (b) $\hat{f}^P$, $k = 50$ (c) $\hat{f}^P - \hat{f}^D \leq 10\%$

Figure 1: Comparison of the kernel density function with an approximation, which is based on k centroids ($k = 50$) for a $2D$ Example.

The remainder of the paper is organized as follows: In section 2, we give an overview of analysis methods for microarray data sets as well as of other related approaches. In section 3, some preliminary work is presented. Our two-phase clustering strategy is then described in section 4. The evaluation is done in section 5. Finally, in section 6, we conclude the paper.

2 Related Work

An overview of state-of-the-art approaches to cluster microarray data sets is given by Chipman et al. in [3]. Using clustering methods, it is possible to identify groups of similar samples (genes). The basis for this is often a similarity measure between genes or samples as a function of the rows or columns in the gene expression matrix. A simultaneous clustering, also called "biclustering," of both genes and samples is proposed in [2].

A similar approach to our two-phase clustering strategy is the Distributed Data Mining (DDM). An overview of some state-of-the-art research results is given in [10]. Januzaj et al. [8] propose a density-based distributed clustering approach. Their recursive technique consists of four different steps, whereas they assume that the data is horizontally distributed. In the first step, the data is clustered locally using the DBSCAN algorithm [5] followed by the determination of local models. In the models, each local cluster is represented by a set of specific core points. These local models are used to compute the global clustering model with the help of DBSCAN, too. In the fourth step, the global result is sent back to the local sites to update the local clustering. This step is necessary to consider data dependencies between local sites. For vertical distributed data sets, a collective hierarchical clustering algorithm is proposed by Johnson et al. [9].

Zeng et al. [15] have developed an adaptive meta-clustering approach combining the information from different clustering results. In their proposal, different clustering results are computed from one single input data set using different algorithms. The objective of their research is to provide a better understanding of the data, because all available clustering approaches are heuristic and can only derive an approximation of the optimal result.

3 Preliminaries

A powerful and effective method to estimate an unknown density function f in a non-parametric way from a set of data points is kernel density estimation [13, 14] and is defined as follows:

Definition 1 (*Kernel Density Estimation*) Let $D \subset \mathbb{R}^d$ be a data set, h be the smoothness level. Then, the kernel density estimate $\hat{f^D}(x)$ based on the kernel K is defined as:

$$\hat{f^D}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

with $0 \leq \hat{f^D}$ and $\int K(x)dx = 1$.

Various kernels K such as the gaussian kernel have been proposed in related literature. The computation of this standard kernel density function requires a number of operations in $O(n)$ (where n is the number of data points) to determine the density at a single point $x \in D$. In recent research activities, much attention has been paid to the development of more efficient density estimation methods.

The WARPing (Weighted Averaging of Rounded Points) framework describes an efficient way to develop density estimation methods based on pre-binning. Zhang et al. [17] propose a density function based on k centroids approximating the standard density function using this framework. Let $C = \{\mu_i \in D : 1 \leq i \leq k\}$ be the set of centroids, $I(x) = min\{i : dist(x, \mu_i) \leq dist(x, \mu_j) \forall j \in \{1, \ldots, k\}\}$ will be the index function delivering the minimal index of the nearest centroid $\mu_i(D) = \{x \in D : I(x) = i\}$ for the set of data points, which have μ_i as the nearest centroid, $n_i = \#\mu_i(D)$ the number and σ_i the standard deviation of the data points in $\mu_i(D)$. The determination of the centroids can be conducted using several vector quantization methods, e.g. k-means, centroid linkage or its popular variant BIRCH [16].

This density estimation based on k centroids uses the sufficient statistics (average, variance and the number of data points) of the Voronoi cells, which are given by the positions of the k centroids. The density function based on the Voronoi prebinning and the gaussian kernel for a given smoothing level h is:

$$\hat{f^C}(x) = \frac{1}{n} \sum_{i=1}^{k} \frac{n_i}{\sqrt{2\pi}\sqrt{\sigma_i^2 + h^2}^d} \cdot exp\left(-\frac{(x - \mu_i)^2}{2(\sigma_i^2 + h^2)}\right) \quad (1)$$

The computation of this density estimation function requires a number of operations in $O(k)$ (where k is the number of cluster centroids) to determine the density at a single point $x \in D$. This method reduces the runtime complexity significantly if $k \ll n$, where n is the number of data points in D. Figure 1 illustrates the impact of the data reduction and the loss of accuracy for a two-dimensional data sets. The density estimation was done with 50 centroids (Figure 1(b)). The loss of accuracy is depicted in Figure 1(c).

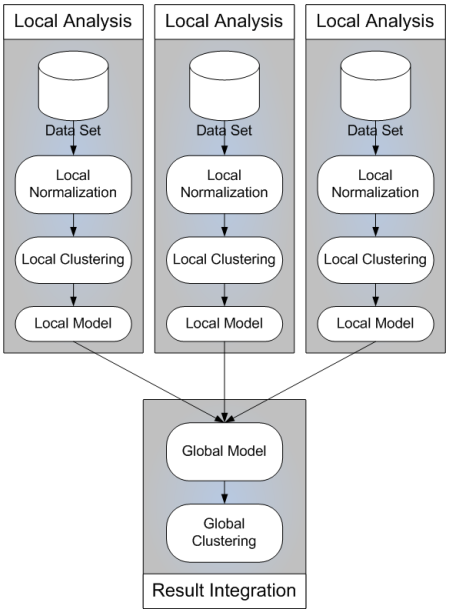


Figure 2: Two-phase Clustering Strategy

4 Two-Phase Clustering Strategy

With our new approach of retrieving cluster information from a set of gene expression data, we propose a novel way to incorporate data from several independent studies. In comparison to the classic approach the structure of the analysis process has changed. Normalization and a first statistical analysis, in our case clustering, are performed for each data set separately (Figure 2). This allows the adjustment of normalization and clustering according to the specific characteristic of the underlying data. Furthermore, data from different studies are grouped based on meaningful statistical values instead of measured raw intensity or ratio values.

To obtain a global interpretable view of l different gene expression data sets $S_1, \ldots, S_l$, an overall clustering result has to be computed. Therefore a set of local clustering results needs to be combined, which

1. consists of a varying number of genes

   $$\overrightarrow{X_i} = \{x_i^1, \ldots, x_i^{n_i}\}, 1 \leq i \leq l,$$

   where n_i is the number of investigated genes in the i^{th} microarray data set,
2. is divided in a varying number of clusters

   $$\overrightarrow{C_i} = \{c_i^1, \ldots, c_i^{k_i}\}, 1 \leq i \leq l,$$

   where k_i is the number of clusters in the i^{th} local clustering result and
3. has been derived from data of different dimensionality according to the numbers of samples within a single study.

A potential algorithm to combine l different local clustering results in gene expression analysis requires to overcome these problems.

4.1 Local Clustering and Local Model

The microarray data sets S_i are usually n_i x m_i matrices, where n_i is the number of genes and m_i is the number of samples. The entries represent the intensities of genes in samples. The first phase of our two-phase clustering strategy is that each gene expression data set S_i is separately normalized and clustered (called local normalization and local clustering). This allows the separate adjustment of normalization and clustering for each data set instead of having to deal with just one fully integrated data base of m different microarray data sets. The result of each local clustering is vector $\overrightarrow{C} = \{c_1, \ldots, c_k\}$, where c_i is the set of genes belonging to the i^{th} cluster. The number of clusters for each microarray data set can be different.

In order to compute a global clustering from a set of l different local clustering results, we require for each local clustering result a local model which satisfies the following aspects:

- All genes of the underlying microarray data sets must be included in the local model. This is necessary to know which genes are investigated in the experiment.
- The local model allows a good approximation of the similarity between genes according to the underlying intensity values and the clustering result.

A local model could consists of an n x k matrix LM, where n is the number of genes and k is the number of clusters. The entries of the matrix are either 0 or 1; while the value 0 indicates that the gene does not belong to the corresponding cluster, the value 1 indicates that the gene does belong to the cluster. The drawback of this local model is the lack of information on the data distribution of the underlying data set and the resulting inability to approximate the similarity of genes in a satisfying manner. This local model allows only to determine which genes belong to the same cluster.

A more accurate local model is also an n x k matrix LM, where the entries represent the density probability of a gene belongs to a cluster. We know that the local clustering is represented by a vector $\overrightarrow{C} = \{c_1, \ldots, c_k\}$, where c_i is the set of genes belonging to the i^{th} cluster. For each cluster, we can compute the centroid μ_i and the standard deviation σ_i. Moreover, we know that each gene belongs only to one cluster c_i. Using the formula (1) we can compute for each gene x the membership to a specific cluster μ_i based on the density estimation.

$$\hat{f^C}(x|\mu_i) = \frac{n_i}{n\sqrt{2\pi}\sqrt{\sigma_i^2 + h^2}^d} \cdot exp\left(- \frac{(x-\mu_i)^2}{2(\sigma_i^2 + h^2)} \right) \tag{2}$$

Compared to [15], we have reduced the computational complexity from $O(n^2)$ to $O(n \cdot k)$. Subsequently, we can determine a normalized density probability $dp(\mu_i|x)$ that a cluster μ_i includes the gene x.

$$dp(\mu_i|x) = \frac{\hat{f^C}(x|\mu_i)}{\hat{f^C}(x)} \tag{3}$$

For each gene $x_i \in \overrightarrow{X}, 0 \leq i \leq n$ in a gene expression data set the density probability for all centroids forms a vector $\overrightarrow{V_{x_i}} = \{dp_1, \ldots, dp_k\}$, where k denotes the number of local clusters. The vectors $\overrightarrow{V_{x_i}}$ for all genes x_i in the microarray data set S_i represent the local model. The complete procedure to determine the local model for a gene expression matrix is illustrated in algorithm 1.

To summarize, in the first phase of our two-phase clustering strategy, we transform each gene expression matrix

into a local model matrix with regard to the local clustering result. The local model matrix has a size of n x k, where n is the number of genes and k is the number of local clusters. The entries represent a normalized density probability that a gene belongs to a cluster. The advantage of this local model is that it includes information on the underlying data distribution and the local clustering result. The resulting local models can be made accessible over the network instead of the raw microarray data sets. The combination of l different local models to obtain a global clustering is presented in the following subsection.

4.2 Global Model and Global Clustering

The local models replace the raw microarray data sets in our approach as starting point for the global clustering. To obtain a global interpretable view of l user-specified different microarray data sets $S_1, \ldots, S_l$ now, an overall clustering result has to be computed from the local models. The first task in the second phase of our two-phase clustering strategy is to determine a global model of the l different local models. The resulting global model should represent the integrated information of the considered l local models, so that a global clustering could be computed. The last task in the second phase is to determine the global clusters from the global model.

The determination of the global model on the basis of the pure local models is difficult because of the different local clustering results. An important observation with regard to the local models is that the distance between two genes x_i and x_j in a local model is a measure of both genes belonging to the same cluster. Furthermore, the distance is an indicator for the similarity of the two genes in the underlying microarray data set.

$$dist(x_i, x_j) = \sqrt{\sum_{l=1}^{k} (dp_{i,l} - dp_{j,l})^2} : 1 \leq i, j \leq n, i \neq j$$

The entries in the local model are normalized density probability values for the event that the gene is included in the cluster. A high distance between two genes x_i and x_j indicates that the genes belong to different clusters and are therefore dissimilar. A small distance indicates that the genes belong to same cluster and are similar in the microarray data set. Using this distance function, a density-based similarity matrix $M : m_{ij} = dist(x_i, x_j), 1 \leq i, j \leq n, i \neq j$ can be derived for each local model. The resulting similarity matrices for the l local models satisfy all requirements to compute a global model in a subsequent step.

DETERMINING LOCAL MODEL
Required: Microarray Data Set S // gene expression matrix

Double[][] LM // local model matrix
S_{norm} = NORMALIZATION(S)
C = CLUSTERING(S_{norm})
for *all* $x_i \in S_{norm}$ **do**
  for *all* $\mu_j \in C$ **do**
    compute $LM[i][j] = dp(\mu_j|x_i) = \frac{\hat{f^C}(x_i|\mu_j)}{\hat{f^C}(x_i)}$
  end
end
return LM

Algorithm 1: Determining Local Model for a microarray data set S

For each of the m local models in our approach, a distance matrix $M^i, 0 \leq i \leq m$ is calculated. As part of the result integration, these matrices need to be combined resulting in one single distance matrix M^{global}:

$$M_{global} = w_1 \bullet M^1 + w_2 \bullet M^2 + ... + w_m \bullet M^m \quad (4)$$

The computation of the global matrix is similar to [15]. The global similarity matrix M^{global} contains the similarity between every two genes of the microarray data set. The global matrix is obtained by weighted addition of local results. The reason and the determination of the weighting factors will be considered in the next section.

From this relationship, a hierarchical and/or a density-based clustering result can be extracted dividing the full data set into groups of similar objects. In the case of a hierarchical clustering, each cluster $C_1, \ldots, C_n$ contains exactly one data point after initialization. In an iterative process, the entry showing the highest similarity between two clusters is identified and the corresponding clusters are merged while the affected entries in the distance matrix are updated. This iteration proceeds till a threshold defining the maximal cost for merging two clusters is reached. The data set gets separated into clusters. The similarity is defined by a distance measure based on a cost function, e.g. single, average or complete linkage. To increase the robustness towards outliers, we have chosen average-linkage hierarchical clustering.

4.3 Weighting of microarray data sets

Not all microarray experiments have the same quality. This fact should be considered in the computation of the global clustering result. The better the attested quality of local results, the stronger their impact on the global result. Aside from statistical properties of a clustering, such as the within-cluster standard deviation [15] or the silhouette index [11], technical and biological criteria could be used to estimate the weighting factors w_i of the local results within a distributed meta-analysis.

The term *technical criteria* refers to properties of the microarray or the experimental procedure itself, which could be used to utilize a quality measure for microarray data sets:

Existence of technical or biological replicates: A meaningful experimental result can be hardly achieved without replicates. While technical replicates are used to validate the labeling and hybridization process, biological replicates are based on separate RNA extraction from different individual samples. It is clear, that at least 3 replicates are necessary for statistical relevance.

Within-replicate variation: If replicates are available, the variation between replicates should be low to indicate that the experiments have been performed with high standards and that results can be reproduced.

Missing data points: For high-quality analysis, we favor complete data sets where the percentage of data points with no measurement available is low. Nevertheless, it frequently happens that some measurements can not be performed due to experimental problems or self-defined constraints, e.g. the exclusion of negative intensity values from the analysis of Affymetrix data sets.

Quality measures for the ranking of the scanning result after hybridization could contain the following features with a_i being a data point (spot) in an microarray scanning result and k being the total number of data points:

Saturation: Good scan results should contain a low percentage of saturated pixels within a data point. Fully saturated data points have to be excluded from the analysis. An averaged saturation factor can be defined as:
$SAT = \frac{1}{k} \sum_{i=1}^{k} \frac{\text{number of good pixels in data point } a_i}{\text{number of all pixels in data point } a_i}$

Shape Additionally, data points should show a compact round shape, signaling that the experiment has been performed successfully. A shape factor per data point can be described as:
$SHP = \frac{1}{k} \sum_{i=1}^{k} \frac{\text{data point area of } a_i}{\text{perimeter of } a_i}$

Homogeneity: The variation of within-data point pixel intensities should be small indicating that hybridization has been performed with high standards using a chip of high quality. The homogeneity of a data point a_i can be defined by the within-data point variance:
$HOM = \frac{1}{k} \sum_{i=1}^{k} var(a_i)$

Brightness: The ratio between foreground and background indicates the amount of bound biological material. For meaningful results a sufficient amount of biological material should be involved resulting in a high signal to noise ratio (SNR) which is commonly defined by $SNR = P_{Signal}/P_{Noise}$ and especially for microarray scan results as:
$SNR = \frac{\text{average foreground} - \text{average background}}{\text{standard deviation of background}}$

A combined technical criterion c_i^T for a data set corresponding to the distance matrix M_i given that all components are considered equally, can be calculated as:

$$c_i^T = SAT_i + SHP_i + HOM_i + SNR_i \quad (5)$$

Weights w_i^T based on technical quality measures for microarray experiments can then be derived accordingly:

$$w_i^T = \frac{c_i^T}{\sum_{i=1}^{m} c_i^T} \quad (6)$$

On the other hand, *biological criteria* could be used to estimate the quality of microarray data participating in a distributed analysis:

Description of sample attributes: Within gene expression experiments, the sample description is of great importance. However, in many data sets the description is inadequate. Therefore, an exact description could be used to weight the experiments, i.e. in the case of a comparison between tumor samples, we might expect the histological grading of the tissue and the survival time of the patients for performing DC to compare gene expression to the grade of the tumor or to survival time.

RNA quality: Criteria for sample RNA quality should be included since the quality of the input material (RNA) defines the quality of the output (gene expression data). Such criteria might be the $28S/18S$ ratio, which should be above 1.75, or quality values from PCR based approaches. With the use of more sophisticated instruments like the Agilent Bioanalyzer. it is possible to define other criteria for RNA quality.

Quality of the probe sequence: Different chip platforms use different types of probe designs. Spotted arrays usually consist of probes generated from PCR products of EST or other sources for a gene. Oligo arrays contain a different number of oligonucleotide probes for a single gene. In general, the annotation of the probes should contain stable identifiers of the common genomic databases used. Moreover, the exact location of the probe on the genome should be provided since splice variants of the RNA of genes might change the measurements. Also cross matching sequences should provided. If PCR products are used, consideration has to be given to the level of evidence on which the PCR product is annotated, i.e. based on sequencing of the product or based on the available annotation of the source sequence, since roughly 20 percent of the available EST clones are wrongly annotated.

In addition, a ranking of studies based on microarray experiments could be defined analogous to a standard scale similar to the existing evidence levels of medical studies [4]. Together with the technical and biological criteria mentioned above, this could lead to a model for the estimation of the weights of single studies within a distributed parallel meta-analysis.

5 Evaluation

For our evaluation, we generated synthetic three-dimensional data sets consisting of normally distributed natural clusters. Therefore, a priori knowledge of the data points' affiliation to clusters was available. In our initial configuration, it contained 1200 data points in four clusters with 600, 400, 100 and 100 data points. From this initial set, four data sets have been derived by changing the clusters' position in space and the within-cluster variation. To obtain test data similar to real microarray data sets, 10% of the data points in each cluster were arbitrarily moved to different clusters. One of the five sample data sets A, B, C, D, E can be seen in Figure 3 showing the generated clusters.

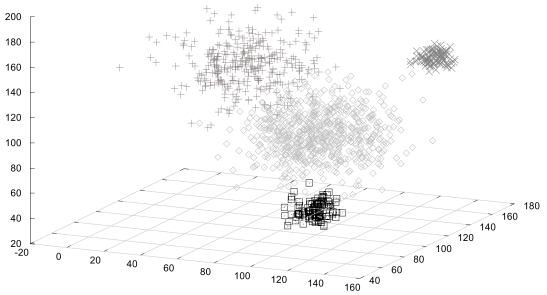


Figure 3: Three-dimensional test data set, labeled as generated.

Following the classic approach the five data sets were integrated, resulting in one multidimensional data set as basis for further data manipulations. Clustering was performed using a hierarchical average linkage algorithm with a Euclidean distance measure. The result is illustrated in Figure 4.

The local clusterings in our two-phase clustering strategy were performed using the k-means or the hierarchical clustering algorithm. The data sets were clustered in 10 clusters, which are more than the expected number of four.

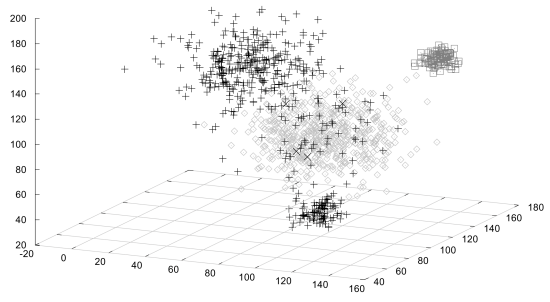


Figure 4: Clustering result for a three-dimensional test data set using the classic approach for meta-analysis.

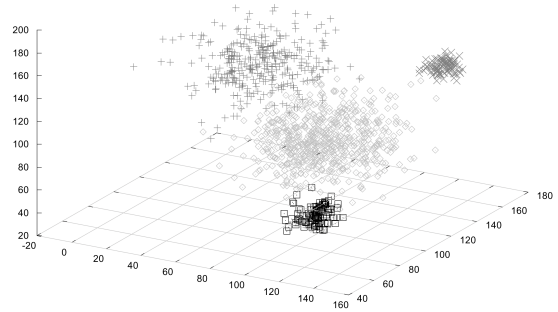


Figure 5: Clustering result for a three-dimensional test data set using our two-phase clustering strategy for meta-analysis.

After result integration, a hierarchical clustering algorithm has separated the data set into four clusters. The result can be seen in Figure 5. The example shows that our two-phase clustering strategy determines more accurate clusters than the classic approach.

In Figure 6, the jaccard indices [1] for the clustering results are shown. We see that for both approaches, the invariant set normalization has separated the data better than the simple scaling has. We also can see that our approach achieved better clustering than the classic approach. The example in Figure 4 illustrates that the classic method tend to assign points from clearly distinguishable clusters to the same cluster, whereas the parallel approach leads to results close to the original cluster structure. As there is no a priory knowledge of the correct clustering for real data, a technical evaluation of the clustering results cannot be performed.

Currently, we are analyzing real data for pancreatic cancer from Friess et al. [6] and Logsdon et al. [12] to perform a high-level comparison based on differentially expressed genes.

6 Conclusion

In this paper, we presented a novel approach for clustering of microarray data sets. Each data set is analyzed individually, which allows adjusted normalization and clustering. The local clustering is followed by the calculation of a local model based on density estimation. A set of local models is combined to a global model using a linear conjunction of derived density-based similarity matrices. Afterwards, the overall clustering can be computed from the global model. Furthermore, we presented an integrated approach to determine the weights of the microarray data sets within our

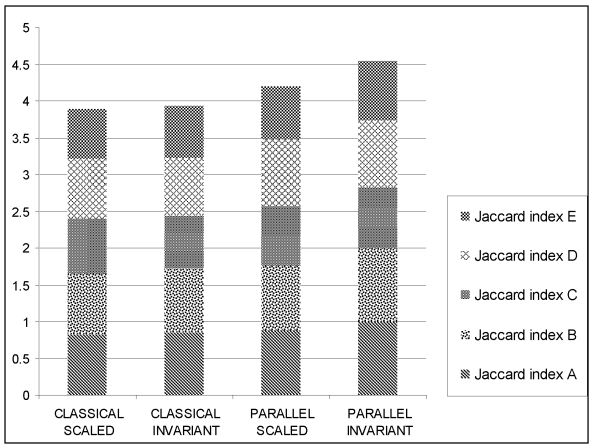


Figure 6: Correct classified pairwise cluster-based occurrence by means of the Jaccard index for test data

method. The weighting factors are determined based on technical as well as biological quality measures. The better the attested quality of local results, the stronger their impact on the global result. The evaluation is done with synthetic three-dimensionally generated data sets consisting of normally distributed natural clusters. Currently, we are analyzing real data for pancreatic cancer from Friess et al. [6] and Logsdon et al. [12] to perform a high-level comparison based on differentially expressed genes.

References

[1] J. Bryan. Problems in gene clustering based on gene expression data. *Journal of Multivariate Analysis*, 90:44–66, 2004.

[2] Yizong Cheng and George M. Church. Biclustering of expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB'00, August 19-23, 2000, La Jolla / San Diego, CA, USA)*, pages 93–103, 2000.

[3] Hugh Chipman, Trevor J. Hastie, and Robert Tibshirani. Clustering microarray data. In Terry Speed, editor, *Statistical Analysis of Gene Expression Microarray Datas*, pages 159–200. Chapman and Hall/CRC, 2003.

[4] Deborah J. Cook, Gordon H. Guyatt, Andreas Laupacis, David L. Sackett, and Robert J. Goldberg. Clinical recommendations using levels of evidence for antithrombotic agents. *Chest*, 108(4 Suppl):227–230, 1995.

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pages 226–231, 1996.

[6] H. Friess, J. Ding, J. Kleeff, L. Fenkell, J. A. Rosinski, A. Guweidhi, J. F. Reidhaar-Olson, M. Korc, J. Hammer, and M. W. Büchler. Microarray-based identification of differentially expressed growth and metastasis-associated genes in pancreatic cancer. *CMLS Cellular and Molecular Life Science*, 60:1180–1199, 2003.

[7] Reinhard Hoffmann, Thomas Seidl, and Martin Dugas. Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biology*, 3:0033.1–0033.11, 2002.

[8] Eshref Januzaj, Hans-Peter Kriegel, and Martin Pfeifle. Dbdc: Density based distributed clustering. In *Proceeding of the 9th International Conference on Extending Database Technology (EDBT'04, Heraklion,Crete, Greece, March 14-18, 2004)*, pages 88–105, 2004.

[9] Erik L. Johnson and Hillol Kargupta. Collective, hierarchical clustering from distributed, heterogeneous data. In *Proceedings of the Large-Scale Parallel Data Mining, Workshop on Large-Scale Parallel KDD Systems, SIGKDD, August 15, 1999, San Diego, CA, USA*, pages 221–244, 1999.

[10] Hillol Kargupta and Philip Chan. *Advances in Distributed and Parallel Knowledge Discovery*. AAAI/MIT Press, 2000.

[11] L. Kaufmann and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley New York, 1990.

[12] Craig D. Logsdon, Diane M. Simeone, Charles Binkley, Thiruvengadam Arumugam, Joel K. Greenson, Thomas J. Giordano, David E. Misek, and Samir Hanash. Molecular profiling of pancreatic adenocarcinoma and chronic pancreatitis identifies multiple genes differentially regulated in pancreatic cancer. *Cancer Research*, 63:2649–2657, 2003.

[13] D.W. Scott. *Multivariate Density Estimation*. Wiley and Sons, 1992.

[14] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 1986.

[15] Yujing Zeng, Jianshan Tang, Javier Garcia-Frias, and Guang R. Gao. An adaptive meta-clustering approach: Combining the information from different clustering results. In *Proceedings of the 1st IEEE Computer Society Bioinformatics Conference (CSB'02, 14-16 August 2002, Stanford, CA, USA)*, pages 276–287, 2002.

[16] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: An efficient data clustering method for very large databases. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada, June 4-6, 1996*, pages 103–114. ACM Press, 1996.

[17] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Fast density estimation using cf-kernel for very large databases. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(San Diego, August 15-18, CA, USA. ACM, 1999*, pages 312–316, 1999.